BJR © 2018 The Authors. Published by the British Institute of Radiology

https://doi.org/10.1259/bjr.20180071

Received: 14 January 2018 | Revised: 04 June 2018 | Accepted: 02 July 2018

Cite this article as:
Demchig D, Mello-Thoms C, Lee WB, Khurelsukh K, Ramish A, Brennan PC. Mammographic detection of breast cancer in a non-screening country. *Br J Radiol* 2018; **91**: 20180071.

FULL PAPER

Mammographic detection of breast cancer in a non-screening country

[1]**DELGERMAA DEMCHIG**, MD, [1]**CLAUDIA MELLO-THOMS**, PhD, [1]**WARWICK B LEE**, PhD, [2]**KHULAN KHURELSUKH**, MD, [3]**ASAI RAMISH**, MD, PhD and [1]**PATRICK C BRENNAN**, PhD

[1]Medical Image Optimization and Perception Group (MIOPeG), Discipline of Medical Radiation Sciences, Faculty of Health Sciences, University of Sydney, Sydney, NSW, Australia
[2]Department of Diagnostic Radiology, Intermed Hospital, Ulaanbaatar, Mongolia,
[3]Department of Diagnostic Radiology, National Cancer Center, Ulaanbaatar, Mongolia

Address correspondence to: Dr Delgermaa Demchig
E-mail: *delgermaa.demchig@sydney.edu.au*

Objective: To compare the diagnostic accuracy between radiologists' from a country with and without breast cancer screening.

Methods: All participating radiologists gave informed consent. A test-set involving 60 mammographic cases (20 cancer and 40 non-cancer) were read by 11 radiologists from a non-screening (NS) country during a workshop in July 2016. 52 radiologists from a screening country read the same test-set at the Royal Australian and New Zealand College of Radiologists' meetings in July 2015. The screening radiologists were classified into two groups: those with less than or equal to 5 years of experience; those with more than 5 years of experience, and each group was compared to the group of NS radiologists. A Kruskal–Wallis test followed by *post-hoc* multiple comparisons test were used to compare measures of diagnostic accuracy among the reader groups.

Results: The diagnostic accuracy of the NS radiologists was significantly lower in terms of sensitivity [mean = 54.0; 95% confidence interval (CI) (40.0–67.0)], location sensitivity [mean = 26.0; 95% CI (16.0–37.0)], receive roperating characteristic area under curve [mean = 73.0; 95% CI (66.5–81.0)] and Jack-knifefree-response receiver operating characteristics figure-of-merit [mean = 45.0; 95% CI (40.0–50.0)] when compared with the less and more experienced screening radiologists, whilst no difference in specificity [mean = 75.0; 95% CI (70.0– 81.0)] was found. No significant differences in all measured diagnostic accuracy were found between the two groups of screening radiologists.

Conclusion: The mammographic performance of a group of radiologists from a country without screening program was suboptimal compared with radiologists from Australia.

Advances in knowledge: Identifying mammographic performance in developing countries is required to optimize breast cancer diagnosis.

INTRODUCTION

The burden of breast cancer remains high in low-income countries. For example, 5year survival rate is approximately 90% in developed countries, but this figure is reported to be around 60% in less developed nations in the Asia Pacific region[1] and late stage at detection is a key contributory agent.[2,3] Although multiple factors are linked to delayed diagnosis, advanced disease presentation suggests that diagnostic efficacy must be improved. Nonetheless, little attention or resources have been directed to this topic, and it is unclear whether it is patient knowledge, disease type, health policy or radiology efficacy that are responsible for disease diagnosis at such a late stage. Until more data around each of these potential causal factors are provided, effective allocation of funding and subsequent successful interventions cannot be initiated.

Currently, mammography is the only screening modality that is proven to reduce mortality from breast cancer, however, many countries still do not have a screening program, which is mostly due to lack of financial resources.[4,5] Nonetheless, opportunistic screening approaches such as private or hospital-based screening have been implemented, yet their diagnostic efficacy is unclear. Interpreting mammograms, especially screening mammograms is a challenging task and prone to diagnostic error, where incorrect, missed or delayed diagnosis are not uncommon.[6–9] Indeed among all imaging modalities, mammography is the most common focus for medical lawsuits against radiologists in the USA.[6,10] Whilst the reasons for this are multifactorial,[11] perceptual and interpretive factors are critically important, and these largely depends on the radiologists' ability to interact effectively with the images being reported.[12] On a positive note, perception and interpretation are modifiable and can be enhanced through appropriate education, training and experience. Nonetheless, diagnostic accuracy varies widely among radiologists,[13–15] and in Australia, diagnostic accuracy varies from 85 to 95% for sensitivity and 80 to 83% for specificity.[7,16] In the non-screening (NS) country where the NS radiologists recruited, these important measures of diagnostic accuracy have never been investigated, and until this first step happens, diagnostic accuracy cannot be optimized.

Australian breast radiologists have demonstrated better accuracy using various parameters when compared to readers from other countries having a screening program.[15,16] This is often explained by differences in reader characteristics such as experience, practice and training. In the NS country, cancer diagnosis relies on diagnostic mammography, with imaging performed following a symptomatic concern. In diagnostic mammography, the expected prevalence of cancer is higher than in screening and presents a very different work process to reading screening mammography, where the expectation is that the very large majority of females will not have cancer.

In this study, we compared radiologists from a NS country to those from a screening country (Australia) using a single population test-set, with the aims of identifying possible differences in diagnostic accuracy and to determine causal factors linking observed measures of diagnostic accuracy with their level of expertise.

METHODS AND MATERIALS

Study participants

All radiologists gave informed consent. The study included a total of 63 breast radiologists. Details on demographic information (age and sex) and professional experience (number of years since qualification as a radiologist; number of cases and hours reading mammography; fellowship training in breast imaging) were self-reported by each reader.

Non-screening radiologists

11 NS radiologists were prospectively recruited and read the cases with no time limit during the BREAST (Breastscreen REader Assessment STrategy),[17] a breast imaging workshop in July 2016 in Mongolia. These radiologists represented approximately 80% of the total number of radiologists involved in mammographic reading in the country, where no specific qualifications are required to report mammograms. The mean age for the NS was 32 years old [standard deviation (SD) = 3.7], the mean number of years reading mammograms was 1.6 years (SD = 1.9) and the mean number of years since being qualified as a radiologist was 4.0 years (SD = 3.6). Weekly reading volume was less than 4 h for all NS radiologists and none of them had completed fellowship training in breast imaging.

Screening radiologists

52 screening radiologists read the cases at the Royal Australian and New Zealand radiologist's meetings in September 2012 and July 2015. Since all our NS radiologists (see above) had less than 5 years' experience reading mammograms, the screen readers were divided into two groups: up to 5 years of experience (n = 27), and more than 5 years of experience (n = 25).

For the less experienced screening radiologists, the mean age was 46 years old (SD = 10.6) and mean number of years since qualifying as a radiologist and working in a screening program were 6.0 years (SD = 7.0) and 2.2 years (SD = 2.0) respectively. 56% had completed fellowship training in breast imaging and 30% spent less than 4 h per week reading mammograms.

The mean age for the more experienced screening cohort was 56 years old (SD = 12) and mean number of years since qualifying as a radiologist and working in a screening program were 19 years (SD = 9.0) and 17 years (SD = 8.7) respectively. 16% had completed fellowship training, and 25% of them spent less than 4 h per week reading mammograms. The characteristics of all radiologists are shown in Table 1.

Test-set

The Breast Screen Digital Imaging Library of New South Wales was the source of all images and all recorded data were deidentified. Each case comprised of craniocaudal and mediolaterial views of the left and right breasts.

The test set included 60 digital mammograms with 20 cancer and 40 normal cases. The images were randomly ordered within the test-set and all participant were exposed with this same order. All 20 cancer cases were biopsy-proven, whilst the normal cases were interpreted as being cancer-free by two screen readers and then had a normal follow-up screening mammograms 2 years later with no interval cancer. Normal images were reported by these readers as being completely normal or containing benign appearances such as oil cysts, intramammary lymph nodes and calcified fibroadenomas. The cancer cases were consisted of four discrete masses, two calcifications, five non-specific densities, five stellate lesions and four spiculated masses. The types of cases selected were representative of those presenting in a breast screening environment.

Test-set reading

All images were read using a pair of five megapixel medical grade display monitors that were calibrated to the Digital Imaging and Communication in Medicine Gray Scale Display Function.

In line with routine radiology practice, each reader was asked to identify and locate lesions suspicious for malignancy or lesions that required further assessment, as well as to provide an interpretation scores in their decision. They were able to mark as many lesions as they could identify on each case using the following interpretation scale: 1, normal; 2, benign; 3, equivocal; 4, suspicious; 5, malignant. If no lesion was marked, the software automatically rated the case as being normal. Readers were not informed of the number of normal and abnormal cases.

Data analysis

Jackknife free-response receiver operating characteristics (JAFROC) figure-of-merit (FOM),[18] receiver operating characteristic, area under curve (ROC AUC), case and location

Table 1. The characteristics of 63 radiologists included in the study

Reader characteristics	Categories	Non-screening radiologists (n = 11)	Screening radiologists with under 5 years' experience (n = 27)	Screening radiologists with over 5 years' experience (n = 25)
Readers' age	≤30	3 (27)	0	0
	31–39	8 (73)	11 (41)	1 (4)
	≥40	0	16 (59)	24 (96)
Number of years since qualification of radiologist	≤5	7 (64)	19 (70)	1 (4)
	6–10	4 (36)	2 (8)	3 (12)
	≥11	0	6 (22)	21 (84)
Number of years reading screening mammogram	0–5	11 (100)	27 (100)	0
	≥6	0	0	25 (100)
Number of cases reading mammograms per week	≤20	10 (91)	7 (26)	0
	21–50	1 (9)	6 (22)	3 (12)
	51–100	0	5 (18)	9 (36)
	101–150	0	4 (15)	3 (12)
	151–200	0	4 (15)	2 (8)
	≥200	0	1 (4)	8 (32)
No of hours reading mammograms per week	≤4	11 (100)	12 (45)	6 (24)
	5–10	0	6 (22)	13 (52)
	11–15	0	2 (7)	3 (12)
	16–20	0	4 (15)	1 (4)
	21–30	0	1 (4)	1 (4)
	≥30	0	2 (7)	1 (4)
Fellowship training in breast imaging	Yes	0	15 (56)	4 (14)
	No	11 (100)	12 (44)	21 (84)

Numbers in parenthesis for reader characteristics are the percentage values.

sensitivity and specificity were calculated for each individual participant. Case sensitivity was defined by the proportion of abnormal cases that were correctly identified as being abnormal. Location sensitivity was defined as the proportion of abnormal cases where the lesion was correctly marked (within 50 pixel of radius from the centre of lesion) and given a confidence score of 3 and above. Specificity was defined by the proportion of normal cases given a confidence score of 2 and below.

ROC is a case-based parameter and thus, when a reader did not locate the cancer correctly but correctly identified the case as abnormal, a correct score is given to that reader. To overcome this issue, a free-response receiver operating characteristics curve was also used.[19] JAFROC is a lesion-based method for analysing free-response data (multiple reader and multiple case studies) and is calculated based on number of true lesion locations and number of normal cases rated by observers. For the observer studies, the JAFROC analysis is useful for assessing reader accuracy in locating cancerous lesions and has greater statistical power than conventional ROC analysis, which neglects location information.[20] JAFROC, therefore, takes into account when some readers correctly identify abnormal cases but incorrectly locate the lesion.

The three groups of radiologists were compared using Kruskal–Wallis test followed by the ranked-based version of Tukey's honest significant difference criterion. MATLAB software 2017 (Mathwork, Natick, MA, USA) was used to perform this statistical test. The individual comparisons were conducted between the following groups:

- NS *vs* screening readers with under 5 years' experience;
- NS *vs* screening readers with over 5 years' experience;
- Screening readers with under 5 years' experience *vs* screening readers with over 5 years' experience

Non-parametric analysis using Spearman's rank order correlation was used to explore the association between diagnostic accuracy and experience parameters of the readers. SPSS software (v. 22) was used for this statistical analysis. A difference with p-value of less than 0.05 was considered as significant.

RESULTS

With regards to reader characteristics, a number of statistical differences were found between the pairs of reader groups and these are shown in Table 2. The NS group was significantly younger than each group of the screen readers, whilst no differences in age were noted between the screening radiologists. The number of hours and number of mammogram

Table 2. Pairwise comparison of reader characteristics for the three groups

Reader characteristics	Non-screening radiologists	Screening radiologists (<5 years' experience)	Screening radiologists (>5 years' experience)	*p*-value
Reader's age	32.0 (29, 35)	43.0 (37, 55)	56.0 (43, 67)	[b]<0.001[a] [c]<0.001[a] [d]0.02[a]
Number of years since qualification as a radiologist	4.0 (1.0, 8.0)	4.0 (1.0, 8.0)	20 (11.5, 22.5)	[b]0.66 [c]0.001[a] [d]0.001[a]
Number of years reading screening mammogram	0.0 (0, 3.0)	1.0 (0, 4.0)	16.0 (10, 22)	[b]0.23 [c]<0.001[a] [d]0.001[a]
Number of cases read per week	1.0[a] (1.0, 1.0)	3.0[b] (1.0, 4.0)	2.0[c] (2.0, 3.0)	[b]<0.001[a] [c]<0.001[a] [d]0.04[a]
Number of hours reading per week	1.0[d] (1.0,1.0)	2.0[e] (1.0,4.0)	2.0[e] (1.5,2.5)	[b]0.003[a] [c]<0.001[a] [d]0.66

[a] a statistically significant difference; (a), Indicates less than 20 cases; (b), Indicates 51–100 cases; (c), Indicates 21–50 cases; (d), Indicates less than 4 hrs; (e) Indicates 5–10 hrs.
[b] Difference between non-screen and screen readers with under 5 years' experience.
[c] Difference between non-screen and screen readers with over 5 years' experience.
[d] Difference between screen readers with under and over 5 years' experience.

cases read per week was also significantly lower for the NS radiologists than it was for each of screening radiologists groups. Correlation analysis, which examined the relationship between the accuracy of all three groups of readers and their reader characteristics, demonstrated no statistically significant associations.

Case sensitivity, specificity, lesion sensitivity, ROC (AUC) and JAFROC (FOM) were calculated for all three groups and summarized in Table 3. For each of performance variables, Table 3 lists the observed means, the variable SDs and the lower and upper bounds for the calculated confidence intervals (CIs). The CIs were calculated using *t*-distribution, sample SD as an estimate for the population SD, σ and confidence level c = 0.95. The calculated mean scores for all performance metrics of the NS radiologists were lower than each of the screening groups of radiologists (Table 3). The lowest performance score for the NS radiologists were location sensitivity [*m* = 0.26; SD = 0.16; 95% CI (0.16–0.37)] whilst the highest value was specificity [*m* = 0.75; SD = 0.14; 95% CI (0.7–0.81)].

Kruskal–Wallis test showed that there were significant differences in sensitivity ($p < 0.001$), location sensitivity ($p < 0.001$), ROC AUC ($p < 0.001$) and JAFROC FOM ($p < 0.001$) between the three groups of radiologists, whilst no significant difference was found in specificity ($p = 0.08$). When comparisons in accuracy were carried out between the NS radiologists and the less experienced screening radiologists, the scores for the NS radiologists were significantly lower in terms of case sensitivity ($p < 0.001$), location sensitivity ($p < 0.001$), ROC AUC ($p < 0.01$) and JAFROC FOM ($p < 0.001$), but not for specificity ($p = 0.84$). The NS group also demonstrated significantly lower accuracy in case sensitivity ($p < 0.001$), location sensitivity ($p < 0.001$), ROC AUC ($p < 0.001$) and JAFROC FOM ($p < 0.001$) compared with more experienced screening radiologists but no difference found in specificity ($p = 0.96$). When two groups of screening radiologists were compared there were no differences in any measured values.

Table 3. Mean performance metrics for the three groups of radiologists

Diagnostic accuracy	Non-screening radiologists		Screening radiologists (<5 years' experience)		Screening radiologists (>5 years' experience)	
	Mean (SD)	95% CI	Mean (SD)	95% CI	Mean (SD)	95% CI
Sensitivity	0.54 (0.2)	0.4–0.66	0.84 (0.1)	0.8–0.88	0.83 (0.15)	0.76–0.89
Specificity	0.75 (0.14)	0.7–0.81	0.76 (0.21)	0.68–0.84	0.81 (0.14)	0.75–0.87
Lesion sensitivity	0.26 (0.16)	0.16–0.37	0.76 (0.14)	0.7–0.81	0.75 (0.15)	0.68–0.81
ROC AUC	0.73 (0.1)	0.66–0.8	0.84 (0.97)	0.8–0.88	0.86 (0.73)	0.83–0.89
JAFROC FOM	0.44 (0.07)	0.4–0.49	0.78 (0.11)	0.73–0.82	0.79 (0.09)	0.75–0.83

CI, confidence interval; JAFROC (FOM), Jack-knife free-response receiver operating characteristics (figure-of-merit); ROC (AUC), receiver operating characteristics (area under curve); SD, standard deviation.
Numbers in parenthesis are the standard deviation (SD).

DISCUSSION

The process of image reporting is a vital component of patient management and it depends largely on individual radiologist's accuracy. The study has compared screening and NS radiologists in screening mammogram interpretation and indicates that radiologists from a NS country have lower sensitivity and accuracy than screening radiologists. It has not addressed the impact of training NS radiologists on performance. We found that the NS radiologists had lesion sensitivity and JAFROC FOM scores that were two and four times lower, respectively, than the less experienced group of screening radiologists. With a case sensitivity of 54% for the NS readers, it is not merely the localization of the abnormality that was a challenge, but also the ability to recognize an abnormal image.

JAFROC FOM presents the probability that a lesion will be given a higher rating score compared with a no-containing lesion location on a normal image. This requires a different understanding of the resultant AUC values as unlike a 0.5 value representing a random score for a normal ROC analysis, this is no longer the case with JAFROC. JAFROC does, however, take into account location information which is a powerful indicator of radiologists' accuracy and is, therefore, a valid and effective measure for assessing diagnostic accuracy of radiologists. In our cases, the JAFROC (0.44) scores for the NS radiologists is impacted by their number of normal images identified correctly, since the specificity was high (75%) whilst the lesion sensitivity was low (26%).

To understand this low performance better, it is worthwhile going back to the eye tracking studies,[12,21] which explained how experts in radiology use an initial holistic review to identify areas of abnormality, followed by visual fixation on the abnormality in question. However, the fact that case sensitivity and lesion sensitivity for the NS radiologists was low may suggest that initial visual detection may be at least part of the reason why these readers missed so many cancers. Such detection relies on having a firm understanding of what constitutes a normal image, so that any abnormal features trigger a rapid response. This requires effective initial training coupled with substantial and ongoing experience. However, in Mongolia, it is difficult for radiologists to establish adequate practice and skills in mammographic interpretation under the current 2-year program in radiology residency, which does not include subspecialty training in breast imaging. In addition, workload in mammography units in Mongolia is relatively low because of the lack of a national screening program.

The ability to recognize normal images is also an important measure of accuracy and previous work[9] has shown that when sensitivity is the same, specificity can be a powerful discriminator between expert and less-experienced observers. Interestingly, no significant differences were found for specificity between any of our groups of readers. Due to the low prevalence of breast cancer,[22] the NS radiologists may have a higher decision threshold when declaring the presence of a cancer, thus potentially explaining both the specificity (the large majority of cases are not expected to have cancer) and sensitivity results (abnormal cases have to be obvious to be declared as abnormal). This hypothesis is supported by number of studies,[23–26] which suggest that *infrequent* targets are often missed, which may explain the overall accuracy of the NS radiologists. This prevalence effect may be ameliorated by exposure to an increasing number of abnormal images,[23] suggesting that tailored educational programs with a range of cancer types may be necessary, since clinical experience may not meet the demands for adequate practice in mammographic examination in Mongolia.

The underlying differences in reader characteristics may explain in part the observed findings on diagnostic accuracy. There was, *e.g.* no evidence of fellowship training amongst the NS radiologists, in contrast to the approximately 45 and 16% of less and more experienced screening radiologists, who had completed such a program. In addition, higher number of readings per year have been shown to be critically linked to better accuracy,[9,15,27] but achieving such experience through clinical practice is unrealistic in Mongolia due to lack of a screening program. We did not find any significant relationships for any of the groups between the reader parameters and measures of diagnostic accuracy, although it has been showed in prior studies that radiologists' experience, practice volume[9,15,28] and attendance of specialized training[13,29] are associated with better accuracy. The absence of this finding may be the result of limited power[30] due to small number of individuals per category, as opposed to representing an acceptance of the null hypothesis.

It should be acknowledged that the test-set originated from Australian females undergoing screening mammography, and therefore, the diagnostic accuracy of the NS radiologists may in part result from the lack of experience dealing with images from such a cohort of females. Ethnically-dependent variations of breast morphology and density could have contributed to our findings. In addition, this study did not involve NS readers who have more than 5 years' experience in mammographic interpretation; however, this reflects the very limited number of experienced breast radiologists in the country. We should finally acknowledge possible screening group familiarity with BREAST test-sets which may have enhanced diagnostic accuracy.

In summary, mammographic performance of the radiologists in a country which does not employ screening is less good at detecting breast cancer in a test-set than radiologists from Australia. The absence of screening program and the associated educational and quality activities is most like a major contributing factor.

ACKNOWLEDGMENT

The student scholarship has been provided by the Australian Government. The authors would like to thank all participating radiologists for their time. We also thank Dr Undrakh-Erdene Erdenebold, Ms Kriscia Tapia and Ms Phuong Trieu for assiting this work.

FUNDING

The work was partially supported by the National Breast Cancer Foundation of Australia (IF-12-02).

REFERENCES

1. Allemani C, Weir HK, Carreira H, Harewood R, Spika D, Wang XS, et al. Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2. *Lancet* 2015; **385**: 977–1010. doi: https://doi.org/10.1016/S0140-6736(14)62038-9
2. Bhoo-Pathy N, Yip CH, Hartman M, Uiterwaal CS, Devi BC, Peeters PH, et al. Breast cancer research in Asia: adopt or adapt Western knowledge? *Eur J Cancer* 2013; **49**: 703–9. doi: https://doi.org/10.1016/j.ejca.2012.09.014
3. Anderson BO, Ilbawi AM, El Saghir NS. Breast cancer in low and middle income countries (LMICs): a shifting tide in global health. *Breast J* 2015; **21**: 111–8. doi: https://doi.org/10.1111/tbj.12357
4. da Costa Vieira RA, Biller G, Uemura G, Ruiz CA, Curado MP. Breast cancer screening in developing countries. *Clinics* 2017; **72**: 244–53. doi: https://doi.org/10.6061/clinics/2017(04)09
5. Teh YC, Tan GH, Taib NA, Rahmat K, Westerhout CJ, Fadzli F, et al. Opportunistic mammography screening provides effective detection rates in a limited resource healthcare system. *BMC Cancer* 2015; **15**: 405. doi: https://doi.org/10.1186/s12885-015-1419-2
6. Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE. Cognitive and system factors contributing to diagnostic errors in radiology. *AJR Am J Roentgenol* 2013; **201**: 611–7. doi: https://doi.org/10.2214/AJR.12.10375
7. Reed WM, Lee WB, Cawson JN, Brennan PC. Malignancy detection in digital mammograms: important reader characteristics and required case numbers. *Acad Radiol* 2010; **17**: 1409–13. doi: https://doi.org/10.1016/j.acra.2010.06.016
8. Reed W, Poulos A, Rickard M, Brennan P. Reader practice in mammography screen reporting in Australia. *J Med Imaging Radiat Oncol* 2009; **53**: 530–7. doi: https://doi.org/10.1111/j.1754-9485.2009.02119.x
9. Rawashdeh MA, Lee WB, Bourne RM, Ryan EA, Pietrzyk MW, Reed WM, et al. Markers of good performance in mammography depend on number of annual readings. *Radiology* 2013; **269**: 61–7. doi: https://doi.org/10.1148/radiol.13122581
10. Goergen S, Schultz T, Deakin A, Runciman W. Investigating errors in medical imaging: lessons for practice from medicolegal closed claims. *J Am Coll Radiol* 2015; **12**: 988–97. doi: https://doi.org/10.1016/j.jacr.2015.03.025
11. Pow RE, Mello-Thoms C, Brennan P. Evaluation of the effect of double reporting on test accuracy in screening and diagnostic imaging studies: A review of the evidence. *J Med Imaging Radiat Oncol* 2016; **60**: 306–14. doi: https://doi.org/10.1111/1754-9485.12450
12. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology* 2007; **242**: 396–402. doi: https://doi.org/10.1148/radiol.2422051997
13. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, et al. Variability in interpretive performance at screening mammography and Radiologists' characteristics associated with accuracy. *Radiology* 2009; **253**: 641–51. doi: https://doi.org/10.1148/radiol.2533082308
14. Miglioretti DL, Ichikawa L, Smith RA, Bassett LW, Feig SA, Monsees B, et al. Criteria for identifying radiologists with acceptable screening mammography interpretive performance on basis of multiple performance measures. *AJR Am J Roentgenol* 2015; **204**: W486–W491. doi: https://doi.org/10.2214/AJR.13.12313
15. Suleiman WI, Lewis SJ, Georgian-Smith D, Evanoff MG, McEntee MF. Number of mammography cases read per year is a strong predictor of sensitivity. *J Med Imaging* 2014; **1**: 015503. doi: https://doi.org/10.1117/1.JMI.1.1.015503
16. Soh BP, Lee WB, Wong J, Sim L, Hillis SL, Tapia KA, et al. Varying performance in mammographic interpretation across two countries: Do results indicate reader or population variances? Proc. SPIE 9787, Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment; 2016.
17. BREAST. *Breast Screen reader assessment strategy*: The University of Sydney; 2011.
18. Chakraborty DP. Recent advances in observer performance methodology: jackknife free-response ROC (JAFROC). *Radiat Prot Dosimetry* 2005; **114**: 26–31. doi: https://doi.org/10.1093/rpd/nch512
19. Chakraborty DP. Clinical relevance of the ROC and free-response paradigms for comparing imaging system efficacies. *Radiat Prot Dosimetry* 2010; **139**: 37–41. doi: https://doi.org/10.1093/rpd/ncq017
20. Chakraborty DP. A brief history of FROC paradigm data analysis. *Acad Radiol* 2013; **20**: 915–9.
21. Kundel HL, Nodine CF, Krupinski EA, Mello-Thoms C. Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. *Acad Radiol* 2008; **15**: 881–6. doi: https://doi.org/10.1016/j.acra.2008.01.023
22. Troisi R, Altantsetseg D, Davaasambuu G, Rich-Edwards J, Davaalkham D, Tretli S, et al. Breast cancer incidence in Mongolia. *Cancer Causes Control* 2012; **23**: 1047–53. doi: https://doi.org/10.1007/s10552-012-9973-2
23. Gur D, Rockette HE, Armfield DR, Blachar A, Bogan JK, Brancatelli G, et al. Prevalence effect in a laboratory environment. *Radiology* 2003; **228**: 10–14. doi: https://doi.org/10.1148/radiol.2281020709
24. Reed WM, Ryan JT, McEntee MF, Evanoff MG, Brennan PC. The effect of abnormality-prevalence expectation on expert observer performance and visual search. *Radiology* 2011; **258**: 938–43. doi: https://doi.org/10.1148/radiol.10101090
25. Evans KK, Tambouret RH, Evered A, Wilbur DC, Wolfe JM. Prevalence of abnormalities influences cytologists' error rates in screening for cervical cancer. *Arch Pathol Lab Med* 2011; **135**: 1557–60. doi: https://doi.org/10.5858/arpa.2010-0739-OA
26. Wolfe JM, Horowitz TS, Kenner NM. Rare items often missed in visual searches. *Nature* 2005; **435**: 439–40. doi: https://doi.org/10.1038/435439a
27. Esserman L, Cowley H, Eberle C, Kirkpatrick A, Chang S, Berbaum K, et al. Improving the accuracy of mammography: volume and outcome relationships. *J Natl Cancer Inst* 2002; **94**: 369–75. doi: https://doi.org/10.1093/jnci/94.5.369
28. Rickard M, Taylor R, Page A, Estoesta J. Cancer detection and mammogram volume of radiologists in a population-based screening programme. *Breast* 2006; **15**: 39–43. doi: https://doi.org/10.1016/j.breast.2005.04.005
29. Miglioretti DL, Gard CC, Carney PA, Onega TL, Buist DS, Sickles EA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. *Radiology* 2009; **253**: 632–40. doi: https://doi.org/10.1148/radiol.2533090070
30. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 2013; **14**: 365–76. doi: https://doi.org/10.1038/nrn3475